Data Preparation For Data Science:
Womit Sie „DATA=“ in den analytischen Procedures von SAS am besten füttern? – Teil 1

Gerhard Svolba

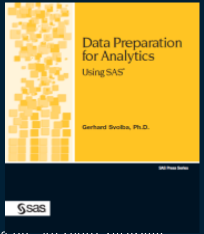
Data Scientist, SAS Austria

#datapreparation4datascience

Medium | LinkedIn | Github | SAS-Books

Youtube: DataPreparation4DataScience

Data Science Use Cases



Weitere Links

- Name: Webinar „Data Preparation for Data Science" im SAS DACH Youtube Channel
- URL: https://www.youtube.com/playlist?list=PLdMxv2SumIKsqedLBq0t_a2_6d7jZ6Akq
-
- Name: Data Preparation for Analytics Using SAS
- URL: https://github.com/gerhard1050/Data-Preparation-for-Data-Science-Using-SAS/blob/master/README.md
-
- Name: Data Quality for Analytics Using SAS
- URL: https://github.com/gerhard1050/Data-Quality-for-Data-Science-Using-SAS/blob/master/README.md
-
- Name: Applying Data Science – Business Analyses Using SAS
- URL: https://github.com/gerhard1050/Applying-Data-Science-Using-SAS/blob/master/README.md

Ein Thema mit vielen Dimensionen

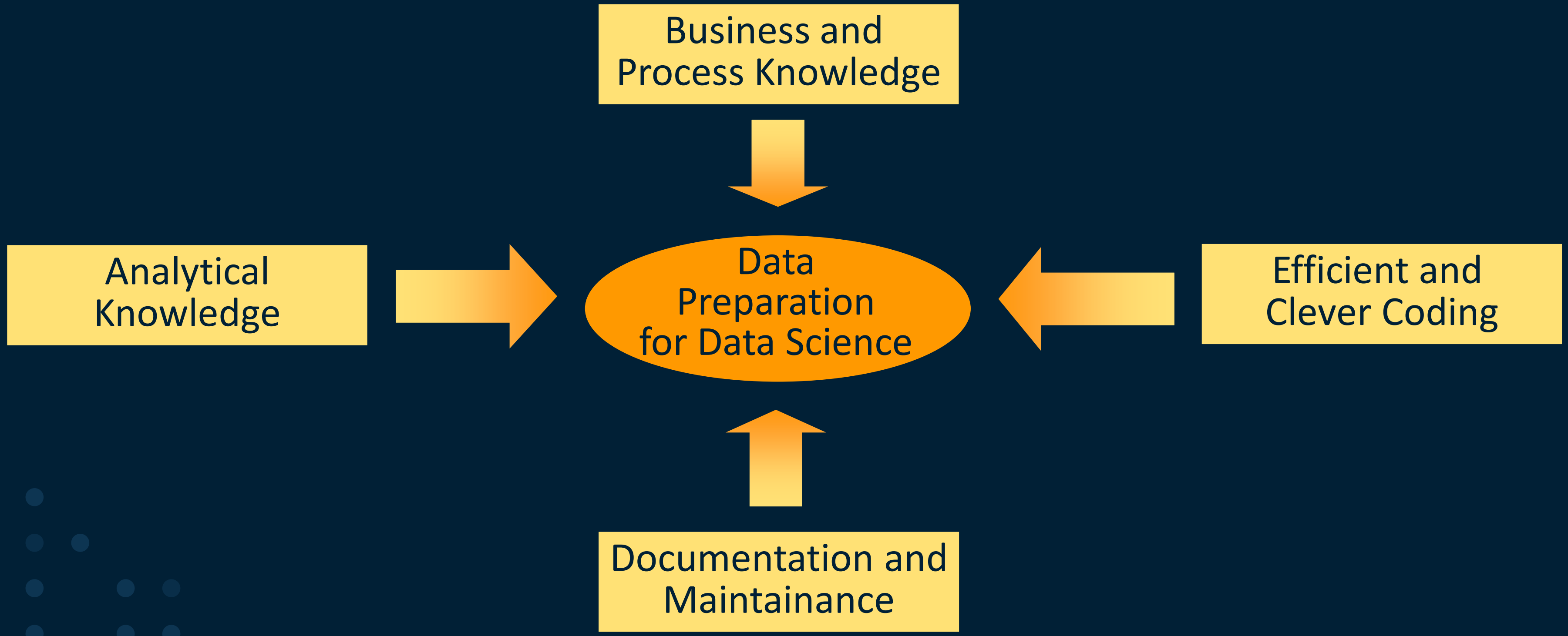
Data Preparation for Data Science

Data Assembly

Data Quality for Analytics

Feature Generation

Four Dimensions
for Data Preparation for Data Science



Data Preparation for Data Science

Data Assembly

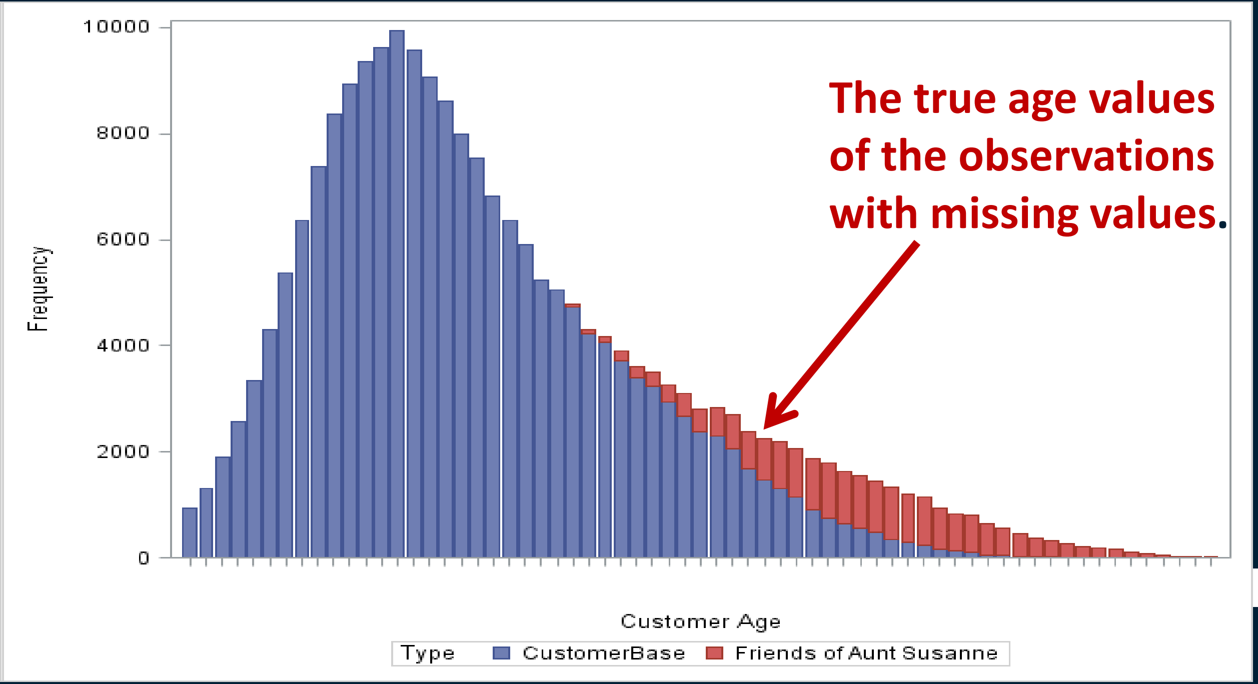
Data Quality for Analytics

Feature Generation

Why my Aunt Susanne gives data scientists a hard time …

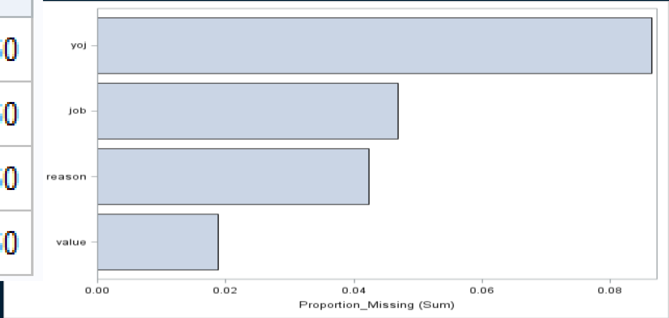
- She got her phone in the mid 1960s.
- Customers‘ „Date of birth“ was of no interest at that time.
- Since the mid 1990s it is mandatory to provide the date of birth on a new contract.
- She never changed her contract type or answered any customer questionnaires.
- She is not the only one with this „data history“.

What does her phone provider see, when he looks at the customer age variable



Typically, missing values are analyzed in a univariate way

Variable	Frequency_Missing	Proportion_Missing	N
YOJ	515	8.64%	5960
JOB	279	4.68%	5960
REASON	252	4.23%	5960
VALUE	112	1.88%	5960



- How many of your variables are infected by the “missing value disease”?
- Not: How many “Full-Records“ do you have?
- Not: Is there a pattern in the structure of missing data?

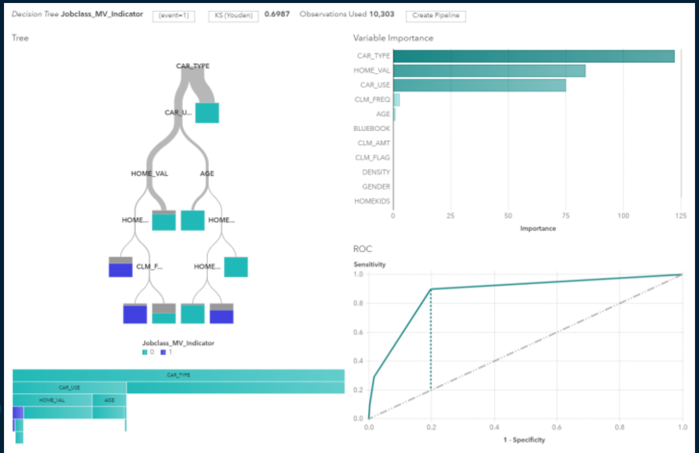
How can you detect and treat such situations?

- Business and process knowledge about the company is key!
- Define imputation rules based on expert knowledge.

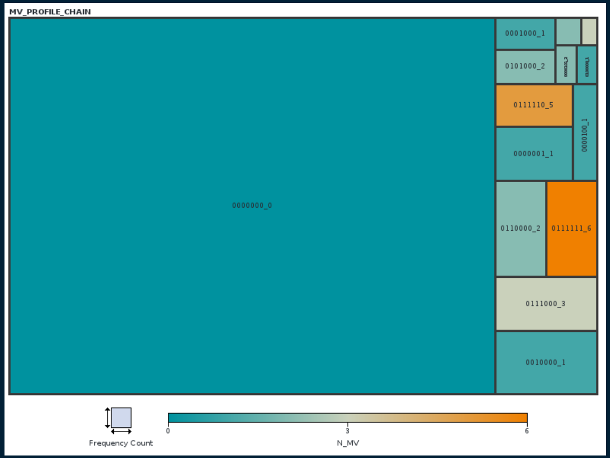
- Simple frequencies per variable do not help.
- Create an indicator variable „Missing YES/NO“ and compare the distribution of other variables like customer start date, product portfolio, …

Get a deeper look into the structure of your missing values

SAS®Visual Analytics provides insight about the nature of your missing values



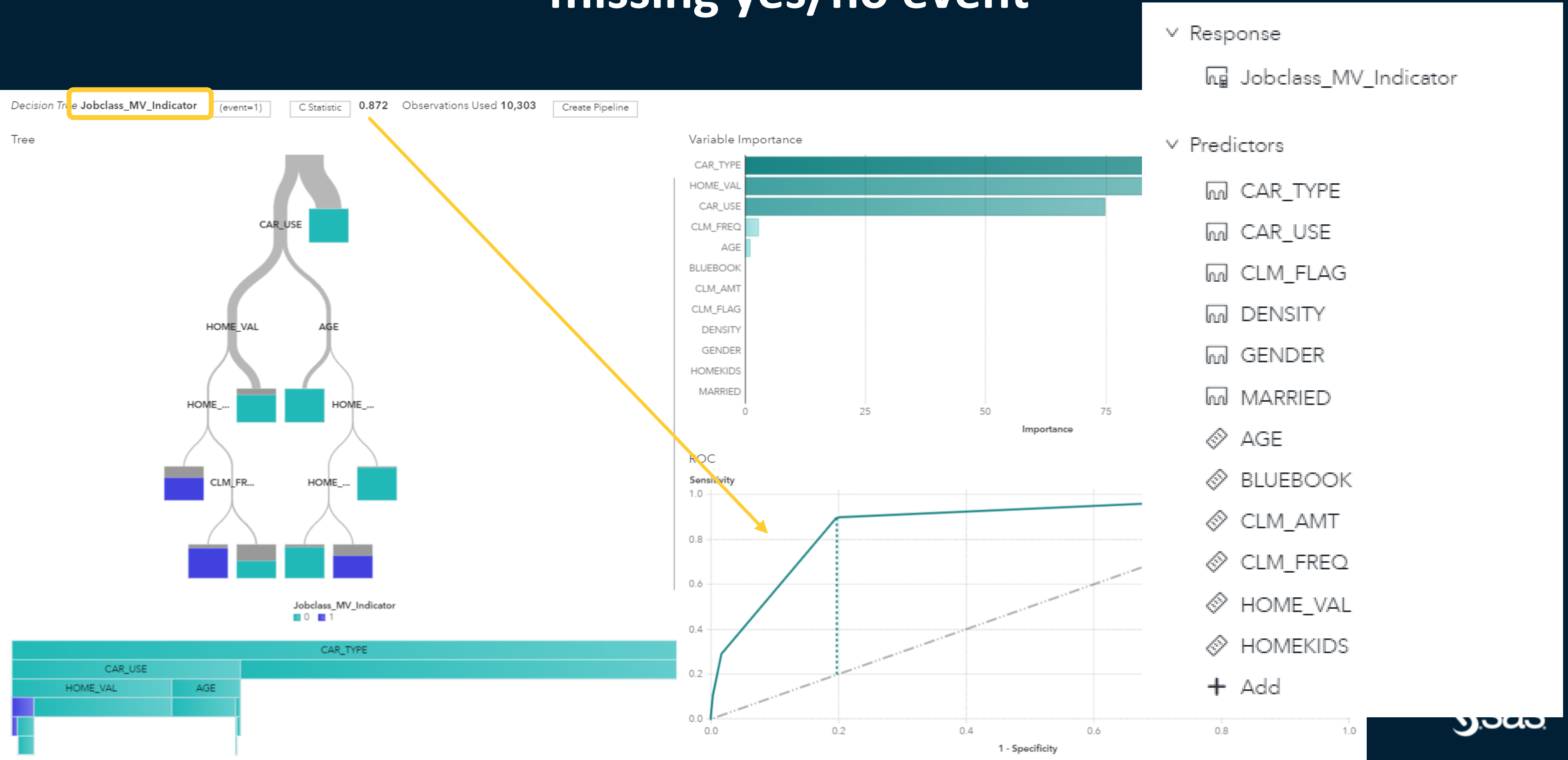
SAS Macro `%MV_PROFILING` to detect pattern in your missing values



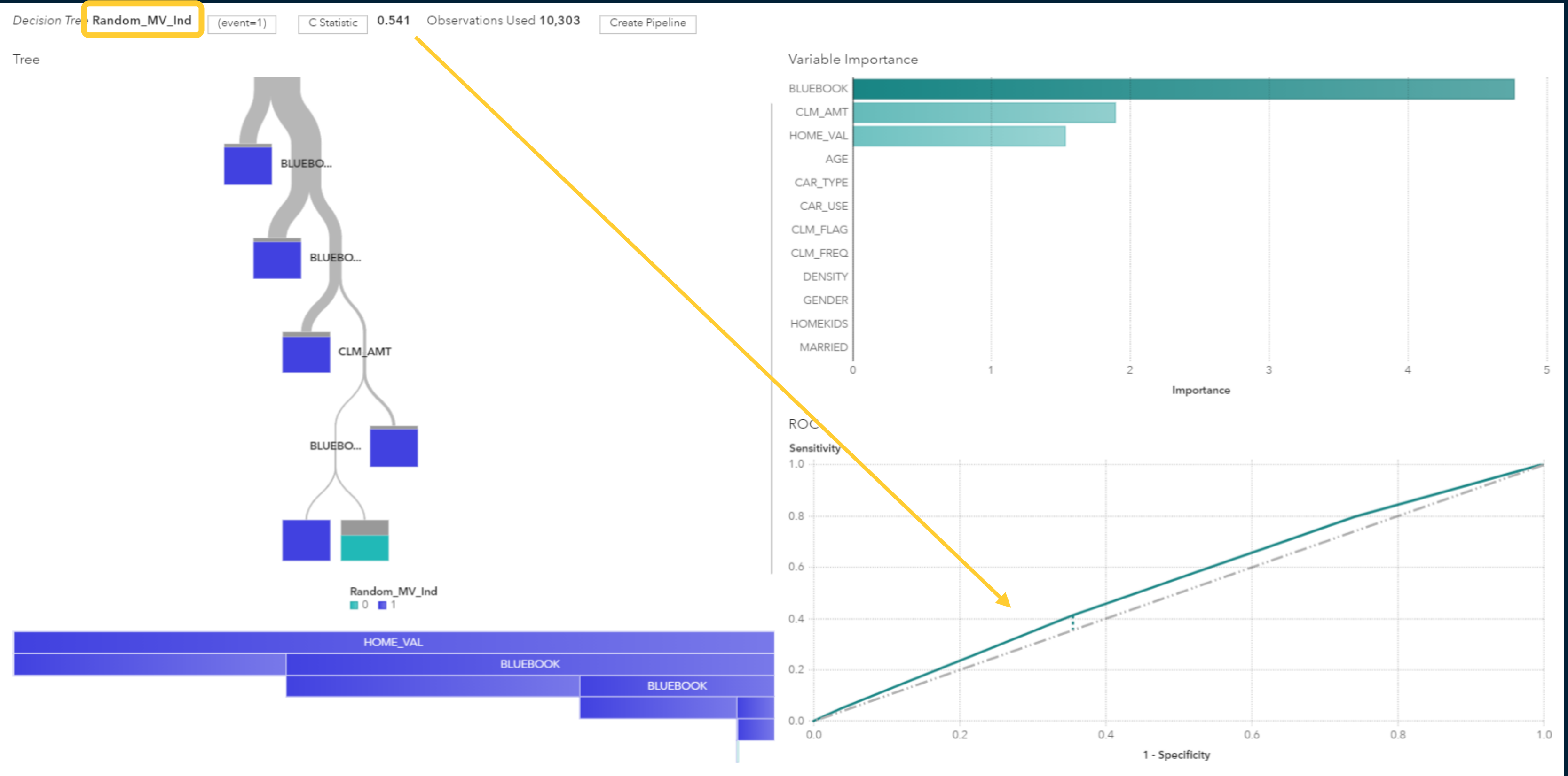
The structure of MISSING VALUES in your data – get a clearer picture with the %MV_PROFILINGmacro

ID	AGE	BLUEBOOK	INCOME	GENDER	JOBCLASS	DENSITY	CAR_TYPE	CLM_FLAG	CLM_AMT
100058542	42	$9,860	.	M	Clerical	Highly Urban	Pickup	Yes	$3,336
100093408	35	$1,500	$4,457	M	Student	Urban	Sedan	Yes	$5,583
100208113	58	$30,460	$102,904	M		Urban	Panel Truck	Yes	$39,104
100237269	45	$16,580	$14,554	F	Student	Rural	SUV	No	$0
10042968	49	$23,030	$99,493	F	Blue Collar	Urban	Pickup	No	$0
100737644	38	$20,730	$95,197	F	Manager	Urban	SUV	No	$0
10078597	60	$27,420	$102,339	F		Highly Urban	Van	Yes	$5,342
100818915	43	$24,360	[illegible]	[illegible]	Lawyer	Highly Ur[illegible]	[illegible]	No	$0
100818915	43	$36,460	[illegible]	[illegible]		Highly [illegible]	[illegible]	No	$0
100818915	43	$20,030	[illegible]	[illegible]		Highly [illegible]	[illegible]	No	$0
100830732	42	$7,520	$7,313	F	Home Maker	Urban	Sports Car	No	$0
100830732	42	$14,300	$7,313	F	Home Maker	Urban	SUV	No	$0
10083678	58	$11,050	$37,528	M		Highly Urban	Pickup	No	$0
100837521	27	$3,770	$27,957	M	Clerical	Highly Rural	Sedan	Yes	$9,117
100896763	38	$15,240	$23,864	M	Clerical	Highly Rural	Sedan	No	$0
101014360	51	$11,560	$58,714	F	Manager	Urban	SUV	No	$0
101209161	76	$29,060	$147,328	F	Blue Collar	Highly Rural	SUV	No	$0
101302659	66	$43,590	$83,827	F	Blue Collar	Urban	Sedan	No	$0
101437485	39	$15,140	$31,869	F	Clerical	Urban	SUV	No	$0
101684050	53	$26,200	$148,193	F	Professional	Urban	Pickup	No	$0
101731472	35	$9,170	$29,250	F	Blue Collar	Highly Urban	SUV	Yes	$4,127
101731472	35	$23,380	$29,250	F	Blue Collar	Highly Urban	Pickup	No	$0
10185862	50	$20,000	$15,989	M	Clerical	Urban	Van	No	$0
10185862	50	$15,330	.	M	Clerical	Urban	Sedan	No	$0
10185862	50	$11,390	$15,989	M	Clerical	Urban	Pickup	No	$0
102077932	26	$20,310	$27,666	F	Clerical	Rural	SUV	No	$0
102080091	29	$6,770	$23,652	F	Clerical	Rural	SUV	No	$0
102262070	46	$21,830	$146,882	M	Manager	Urban	Van	No	$0
102318953	27	$4,200	$35,851	F	Blue Collar	Highly Urban	Sports Car	Yes	$4,102
102411690	52	$6,700	$77,351	M		Urban	Pickup	Yes	$1,070
102503044	41	$13,030	$0	F	Home Maker	Urban	SUV	Yes	$14,509
102778835	56	$14,630	$75,265	M	Manager	Urban	Sedan	No	$0
102861617	62	$13,070	$53,698	F	Professional	Rural	Sports Car	Yes	$5,064
102863719	53	$10,010	$32,073	M	Blue Collar	Rural	Pickup	No	$0
103108106	49	$16,010	$96,143	M	Blue Collar	Highly Rural	Sedan	No	$0
103293644	34	$9,810	$45,384	F	Clerical	Rural	SUV	No	$0
103293644	32	$13,030	$36,179	F	Clerical	Highly Rural	Sports Car	Yes	$4,307
103544547	59	$48,380	$119,537	F	Professional	Urban	Panel Truck	Yes	$6,290
103663917	39	$10,840	$3,414	M	Student	Highly Urban	Sedan	No	$0

Building a decision tree in SAS Visual Analytics to “explain” the “missing yes/no event”



Cross-Check for a randomly generated YES/NO flag



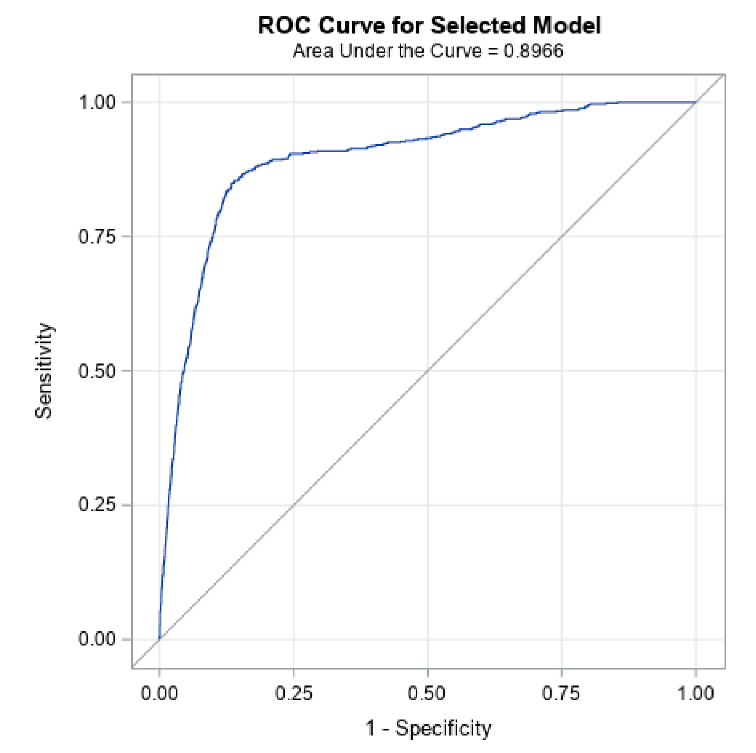
You can also use SAS Procedures for this task

```
data claims_mv;
 set em.claims;
 Jobclass_MV_Indicator = missing(jobclass);
run;


proc logistic data=work.claims_mv plots=roc;
 class car_type car_use clm_flag density gender married;
 model Jobclass_MV_Indicator(event='1')
                    = car_type car_use clm_flag density gender married
                      age bluebook clm amt clm freq home val Homekids
                      /selection=stepwise;
run;
```

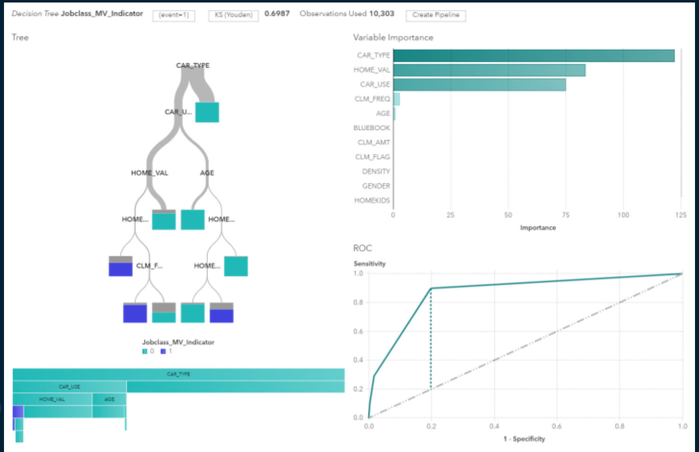
Stepwise Regression "finds" a relevant model → there might be a pattern

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-5.3928	0.3029	316.8831	<.0001
CAR_TYPE	Panel Truck	1	1.3656	0.1398	95.4208	<.0001
CAR_TYPE	Pickup	1	0.7886	0.1186	44.2303	<.0001
CAR_TYPE	SUV	1	-1.0089	0.2012	25.1544	<.0001
CAR_TYPE	Sedan	1	-0.9678	0.1846	27.4970	<.0001
CAR_TYPE	Sports Car	1	-1.3049	0.3502	13.8855	0.0002
CAR_USE	Commercial	1	0.9109	0.0761	143.3348	<.0001
CLM_FLAG	No	1	0.2057	0.0579	12.6161	0.0004
DENSITY	Highly Rural	1	-2.3337	0.7573	9.4965	0.0021
DENSITY	Highly Urban	1	1.2405	0.2659	21.7688	<.0001
DENSITY	Rural	1	-0.2357	0.2962	0.6330	0.4263
MARRIED	No	1	0.3556	0.0539	43.5210	<.0001
BLUEBOOK		1	0.000019	6.856E-6	7.9859	0.0047
HOME_VAL		1	3.179E-6	3.674E-7	74.8721	<.0001

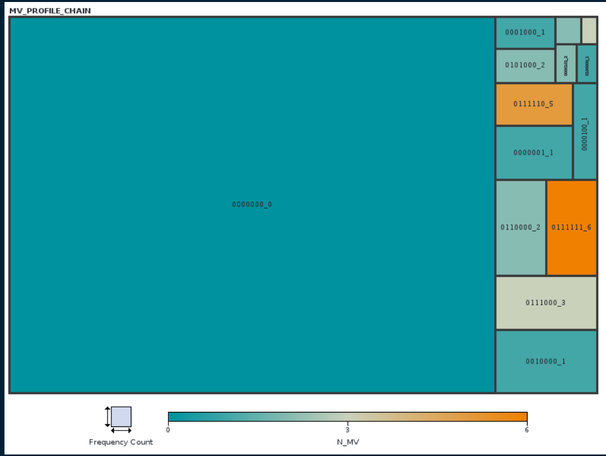


Get a deeper look into the structure of your missing values

SAS®Visual Analytics provides insight about the nature of your missing values



SAS Macro `%MV_PROFILING` to detect pattern in your missing values



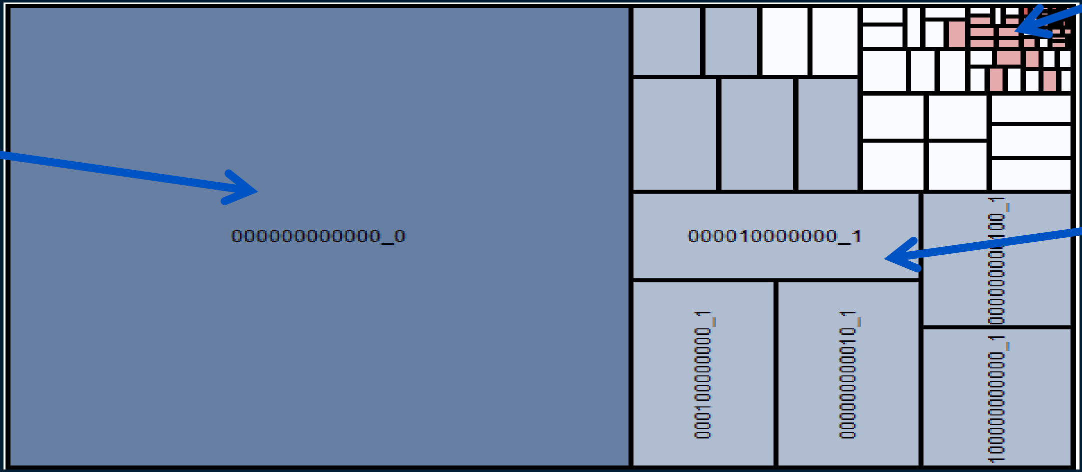
The structure of MISSING VALUES in your data – get a clearer picture with the [%MV_PROFILING]macro

Profiling the pattern of missing values with the %MV_PROFILING macro

- Concatenate each “Missing-Value” Indicator to a string. E.g: 00100100

Records with > 4 missing values

Full Records

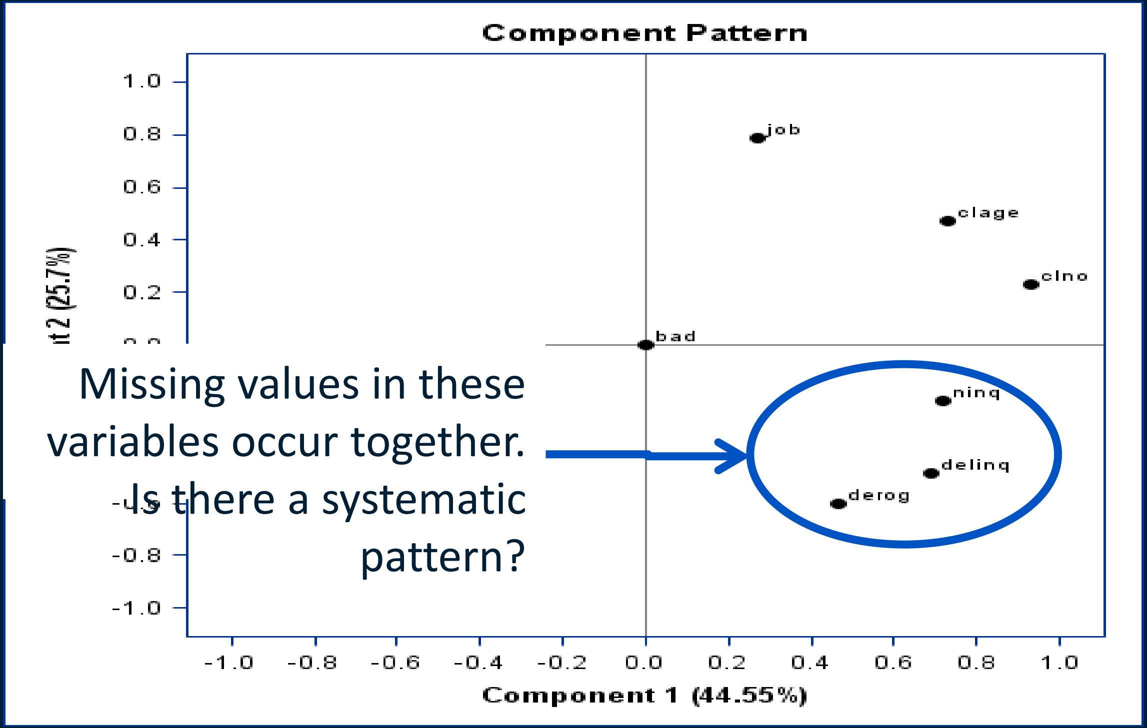


Records with a missing value in one variable

- Macros can be downloaded from #github

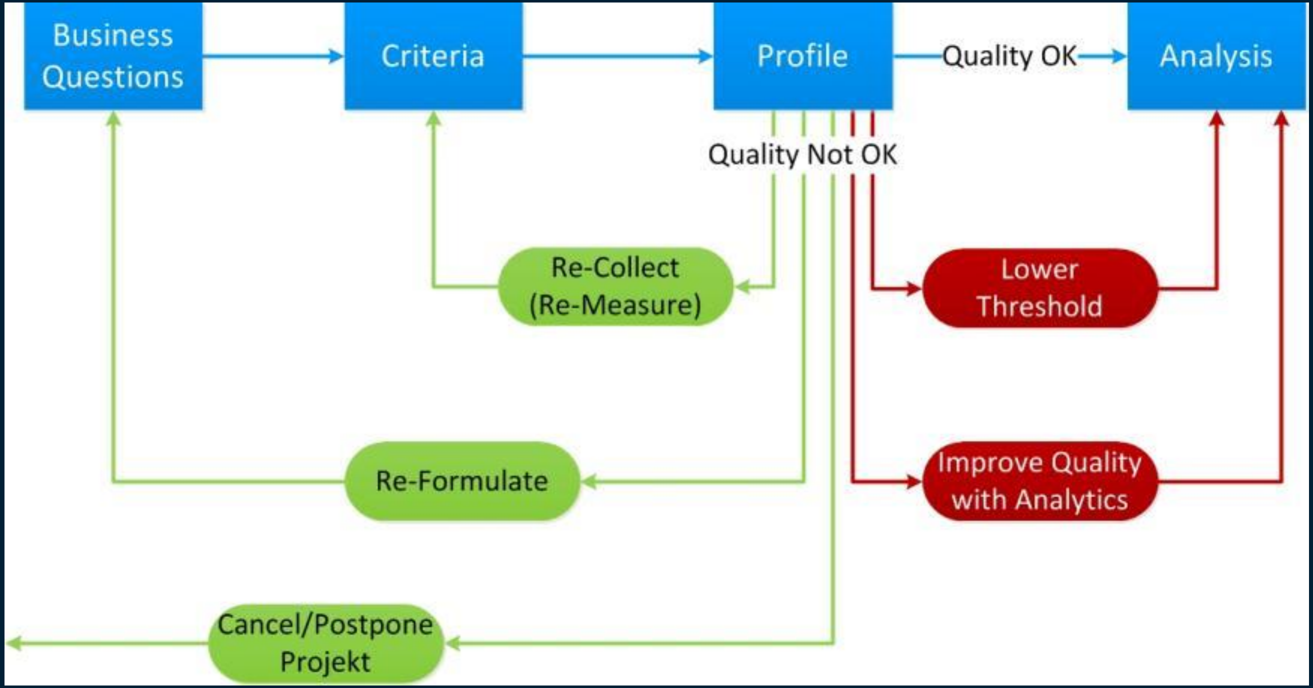
sas

Multivariate analysis uncovers systematic patterns



Missing values in these variables occur together. Is there a systematic pattern?

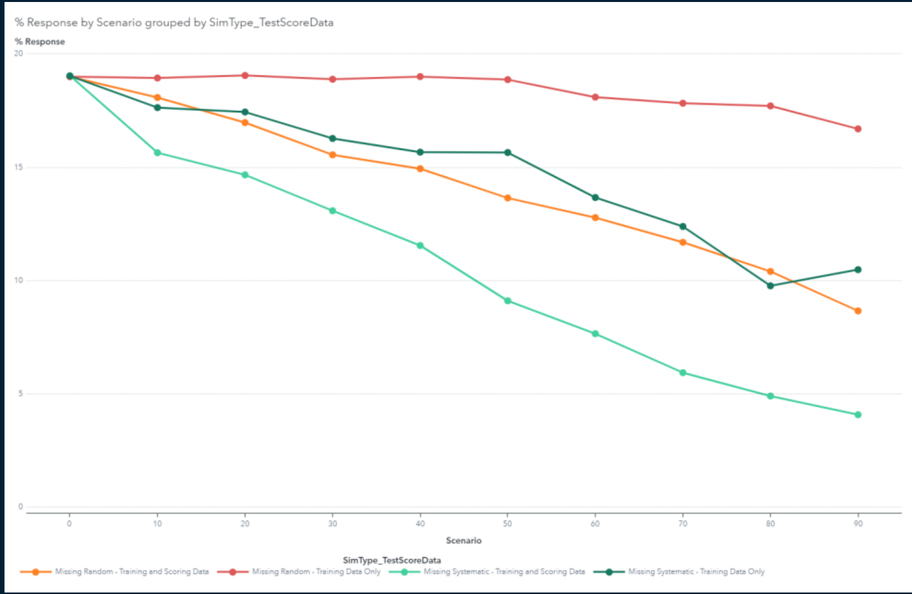
These are your options, if you learn that data quality is poor



Cost
Time, Delays
No Results

Trust
Risk of wrong decisions
Insignificance

Results from simulation studies for the effect of bad data quality on model accracy



- Random missing values in training data only have limited effect
- Missing values that occur also in the scoring data have a larger effect
- Systematic missing values have a much larger effect in general
- Takeaway:
  Not only discuss the “acceptable percentage of missing values” in your data.
  Discuss the WHY they are missing and whether this also occurs in scoring.

Data Preparation for Data Science

Data Assembly

Data Quality for Analytics

Feature Generation

Main Types of Analytic Data Sturctures

One-Row-per-Subject Data Mart

	Customer ID	Date of Birth	Age (years)	Gender	Marital Status	Academic Title	Has Title? 0/1	Branch Name	Customer Start Date	Customer Duration (months)
1	1000002	26DEC1958	44	Male	Married		0	Fil1	01JAN2000	41
2	1000005	25JUN1947	56	Male	Single	Ing.	1	Fil4	01APR1999	50
3	1000006	10DEC1945	57	Female	Married		0	Fil4	01SEP1996	81
4	1000007	02JUN1934	69	Male	Married		0	Fil1	01SEP1997	69
5	1000008	15DEC1957	45	Male	Single	Dr.	1	Fil3	01JAN1996	89
6	1000009	11MAR1959	44	Male	Single		0	Fil2	01JUL2001	23
7	1000014	23AUG1952	51	Male	Single		0	Fil4	01MAY1996	85
8	1000015	12MAY1959	44	Male	Single		0	Fil2	01FEB1999	52
9	1000016	11FEB1967	36	Male	Married		0	Fil2	01FEB2001	28

Multiple-Row-per-Subject Data Mart

	CUSTOMER	TIME	PRODUCT
1	0	0	hering
2	0	1	comed_b
3	0	2	olives
4	0	3	ham
5	0	4	turkey
6	0	5	bourbon
7	0	6	ice_crea
8	1	0	baguette
9	1	1	soda
10	1	2	hering
11	1	3	cracker
12	1	4	heineken
13	1	5	olives
14	1	6	comed_b
15	2	0	avocado
16	2	1	cracker
17	2	2	artichok
18	2	3	heineken
19	2	4	ham
20	2	5	turkey
21	2	6	sardines

Longitudinal Data Mart

	Date	ELECTRO	GARDENING	TOOLS
1	15/08/05	15725	13913	9441
2	16/08/05	15120	16315	9922
3	17/08/05	16631	18996	11345
4	19/08/05	18080	16325	9326
5	20/08/05	15604	14690	9108
6	21/08/05	14518	14388	9371
7	22/08/05	13048	15249	8390
8	23/08/05	13857	13974	10982
9	24/08/05	14869	15704	12104
10	26/08/05	12262	13836	8112
11	27/08/05	15011	13438	8599
12	28/08/05	13612	12625	8389
13	29/08/05	11546	13566	8249
14	30/08/05	21352	16918	13337
15	31/08/05	22900	20813	14099
16	02/09/05	15333	15626	8896
17	03/09/05	13156	13306	8082
18	04/09/05	19294	16361	16267
19	05/09/05	15917	15587	15539

Transposing Data between One-Row-Per-Subject and Multiple-Row-Per-Subject

	ID	TIME	WEIGHT
1	1	1	77
2	1	2	79
3	1	3	83
4	2	1	62
5	2	2	58
6	2	3	59
7	3	1	99
8	3	2	97
9	3	3	92

Makewide →

← Makelong

	ID	weight1	weight2	weight3
1	1	77	79	83
2	2	62	58	59
3	3	99	97	92

Values or “Behaviour” ?

- Location
- Trend
- Pre-Post-Differences
- Variability
- …

The One-Row-Per-Subject Paradigm

Analysis Subject Master Table

ID	Birth	Sex	Region	...	...
1					
2					
3					
4					

Multiple Observation per Analysis Subject

ID	Month	Type	Billing	Usage	...
1					
1					
1					
2					
2					
3					
3					
3					
4					
4					
4					
4					

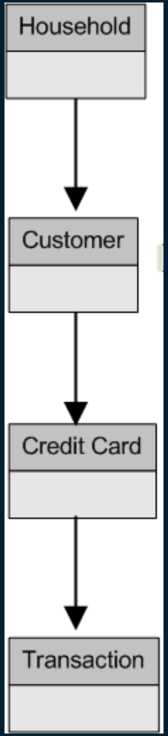
Copy Variables
Create Derived Variables

Aggregate, Transpose Data
Describe Behaviour

Analsis Data Mart

ID	Birth	Sex	Region	Age	...	Billing_Sum	Billing_Mean	Usage_Sum	Usage_Trend	Usage_Variab	N_Trx
1											
2											
3											
4											

Hierarchies:
Aggregating Up + Copying Down



Data Preparation for Data Science

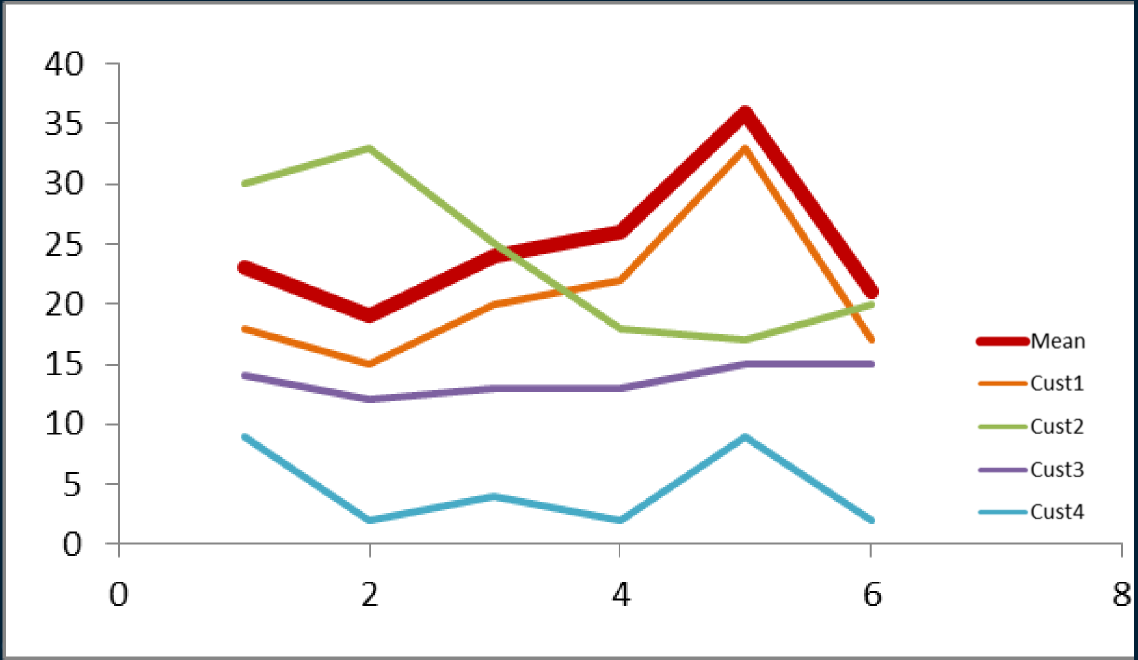
Data Assembly

Data Quality for Analytics

Feature Generation

Describe customer behaviour over time

(displaying 4 example customers and the overall mean)



Cust ID	Level
1	=-
2	=
3	-
4	--

```sas
/*** Step 1      Calculate the usage average
                 per time interval ***/

proc sql;
 create table monthly_average
 as select month,
      mean(usage) as MonthlyAverage format = 8.2
    from usage
    group by month
    order by month;
quit;
```

MonthlyAverage
59.01
62.79
55.11
62.68
54.59
58.49

```
/*** Step 3      Calculate the the correlations between
                 individual value and interval mean ***/

proc corr data = usage_enh
          outp=corr_usage noprint;
 var usage;
 with MonthlyAverage;
 by custid;
run;
```

CustID	_TYPE_	_NAME_	Usage
1000002	MEAN		53.5
1000002	STD		15.732132723
1000002	N		6
1000002	CORR	MonthlyAverage	0.0857280482
1000005	MEAN		35.5
1000005	STD		4.7222875812
1000005	N		6
1000005	CORR	MonthlyAverage	-0.40439214
1000006	MEAN		113.33333333
1000006	STD		7.5277265271
1000006	N		6
1000006	CORR	MonthlyAverage	0.5711078825
1000007	MEAN		49
1000007	STD		1.2247448714
1000007	N		5

```
/*** Step 4      Rearrange to a one-row-per-subject
                 structure ***/

proc transpose data=corr_usage
        out=Customer_ABT(drop=_name_);
  by custid;
  id _type_;
  var usage;
run;
```

1	=-	++	++
2	=	++	--
3	-	=	~
4	--	+	+

CustID	MEAN	STD	N	CORR
1000002	53.50	15.73	6	0.09
1000005	35.50	4.72	6	-0.40
1000006	113.33	7.53	6	0.57
1000007	49.00	1.22	5	-0.50
1000008	38.67	8.50	6	0.03
1000009	31.67	6.19	6	-0.09
1000014	70.83	8.95	6	-0.12
1000015	99.67	8.29	6	0.41
1000016	54.67	3.93	6	0.06
1000018	40.83	27.07	6	0.86
1000019	52.33	12.18	6	0.26
1000021	32.67	3.50	6	-0.47
1000022	115.17	9.70	6	0.61

Feature Engineering – Be creative!

Multiple Observation per Analysis Subject					
ID	Month	Type	Billing	Usage	…
1					
1					
1					
2					
2					
3					
3					
3					
4					
4					
4					
4					

Aggregate, Transpose
Describe Behaviour

Billing_Sum	Billing_Mean	Usage_Sum	Usage_Trend	Usage_Variab	N_Trx

Interval Data

- Correlation of Values
- Course over Time
- Concentration of Values
- Seasonal Pattern

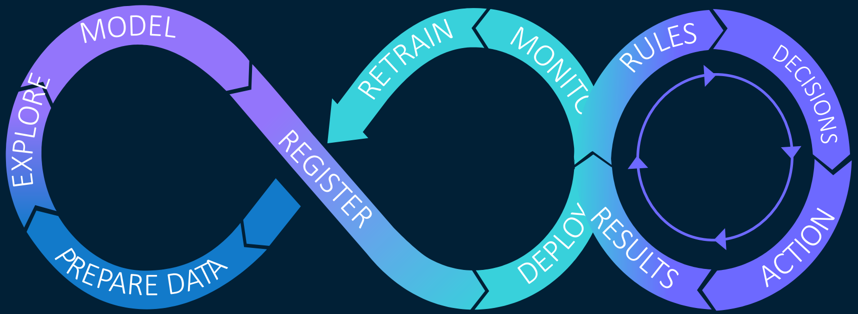
Categorical Data

- Frequency Counts
- Concatenated Frequencies
- Total and Distinct Counts

- Network Data
- Textual Data
- Images and Videos
- …

Conclusion

- Data Preparation is all over the analytic lifecycle!



- Data Preparation is much more than just coding!

All you need to prepare your data for data science is available in the integrated SAS Viya platform

- Data Preparation / Data Quality / Feature Engineering / Variety of Analytical Methods / Visualizing Relationships / Comparing Models / What-If Scenarios / Access for different Persona Roles / Model Ops / ...

Data Preparation for Data Science

Data Assembly | **Data Quality for Analytics** | **Feature Generation**

Gerhard Svolba,
Data Scientist @SAS
mailto: gerhard.svolba@sas.com

[Medium](#) | [LinkedIn](#) | [Github](#) | [SAS-Books](#)
Youtube: [DataPreparation4DataScience](#)
[Data Science Use Cases](#)

Articles and Blogs — Medium, LinkedIn

Webinars — YouTube

Tipps & Tricks — SAS COMMUNITIES

Macros & Downloads